4

The Normal Curve

Earlier I said there were times when it's best to approach the field of statistics without giving much thought to where you're going. This is one of those times. In fact, I'm going to ask you to take a step forward, develop a solid understanding of some information, and do all of it without one thought as to where we're headed. I know that's a lot to ask, but as the phrase goes: Trust me; there's a method to all of this.

We'll begin our discussion where we left off in the last chapter by asking a central question: Why all the fuss about normal curves? As it turns out, scientists long ago noticed that many phenomena are distributed in a normal fashion. In other words, the distributions of many different variables, when plotted as graphs, produce normal curves. Height and weight, for example, are frequently cited as variables that were long ago recognized as being normally distributed.

Having observed that many variables produce a normal distribution or curve, it was only natural that statisticians would focus an increasing amount of attention on normal curves. And so it was that a very special case of a normal curve was eventually formulated. This rather special case eventually came to be known as the **standardized normal curve**.

In one sense, the standardized normal curve is just another normal curve. In another sense, though, it's a very special case of a normal curve—so much so that statisticians often refer to it as the normal curve. Statisticians also use expressions such as *standardized normal distribution*. Regardless of the name—standardized normal curve, the normal curve, or standardized normal distribution—the idea is the same.

As you'll soon discover, the standardized normal curve is a theoretical curve that serves as a basis or model for comparison. It's a point of reference—a standard against which information or data can be judged. In the world of inferential statistics, you'll return to the standardized normal curve time and time again, so a solid understanding is imperative. To better understand this special curve, though, let's start by taking a look at some other normal distributions—ones you might find in the real world.

Before We Begin

Before we get started, let me ask you to think about two concepts. First, I want you to think about the concept of a percentage. Then I want you to think about the concept of a dollar. I know, that may sound very strange, but let me urge you to go along with this. There's a lesson to be learned.

Let's start with the idea of a percentage. Think about what a percentage tells you and how often you rely upon that concept when you communicate. For example, maybe someone tells you that there was a 15% drop in sales at the local grocery store last month. Another person tells you that enrollment at the local college increased by 6%. The use of a percentage to express some amount allows you to conjure up a mental image of a decrease or increase. Because a percentage represents a standard, so to speak, it's often very helpful when you want to make comparisons. For example, let's say your professor tells you that 14% of your class made a score of B, but 22% of the afternoon class made a grade of B. It really doesn't matter how many students are enrolled in each class; the percentage figures allow you to conjure up a mental image about the relative performance of students in the two classes.

In a way, you can think of a dollar in the same terms. To understand this, let me ask you to think about the concept of a dollar, but don't think of a dollar bill that's in your pocket. Instead, think about the notion of dollar as something that you rely upon as a basis for comparison. For example, let's say you've been surfing the net in search of a bargain on a television set. You find two televisions that interest you, but there's a problem. The price of one set (manufactured in Japan) is given in Japanese currency (yen), while the price of the other set (manufactured in Germany) is given in European currency (the euro). Any initial confusion you might experience is quickly erased as you begin to work your way through the situation. It's a simple matter of converting each currency (yen and euro) to dollars. Once you've done that, you're in a position to make a comparison. And that's the point. A dollar, at least in that example, isn't something tangible. Instead, it is something abstract. But a dollar, in an abstract sense, becomes essential to your ability to compare one price to the other.

Although examples about percentages and dollars might strike you as strange, they're relevant to the material that you're about to encounter. They demonstrate the importance of having a means of comparison—some sort of standard or basis that we can use as the foundation for our comparison. And that, in a nutshell, is where we're going in this chapter.

Real-World Normal Curves

Ordinary normal curves—curves like some of the ones we considered in the last chapter—are always tied to empirical or observed data. An example might be a collection of data from a drug rehabilitation program. Let's say, for example, someone gives you some summary information about the amount of time participants spend in voluntary group counseling sessions. Assume that you only know summary information, that you don't have detailed data. Let's also assume you've been told the data are normally distributed, with a mean of 14.25 hours per week and a standard deviation of 2.10 hours.

Because you know that the data reflect a normal distribution, you're in a position to figure out quite a lot, even if you don't have the actual data. For example, using some of the information you learned in the last chapter, you could quickly determine that approximately 68% of the program participants spend between 12.15 and 16.35 hours in voluntary group counseling. To refresh your memory about how you could do that, just follow the logic:

1. You know that the mean is 14.25 hours.
2. You know that the standard deviation is 2.10 hours.
3. You know that the data are distributed normally (the distribution is normal).
4. You know that 68% of the area or cases under a normal curve falls between one standard deviation above and below the mean.
5. Add one standard deviation to the mean to find the upper limit: 14.25 + 2.10 = 16.35 hours.
6. Subtract one standard deviation from the mean to find the lower limit: 14.25 – 2.10 = 12.15 hours.
7. Remembering the important point that the area under the curve really represents cases (program participants, for example), express your result as follows: Approximately 68% of the program participants spend between 12.15 and 16.35 hours per week in voluntary group counseling.

To further grasp the logic of this process, consider the illustration in Figure 4-1.

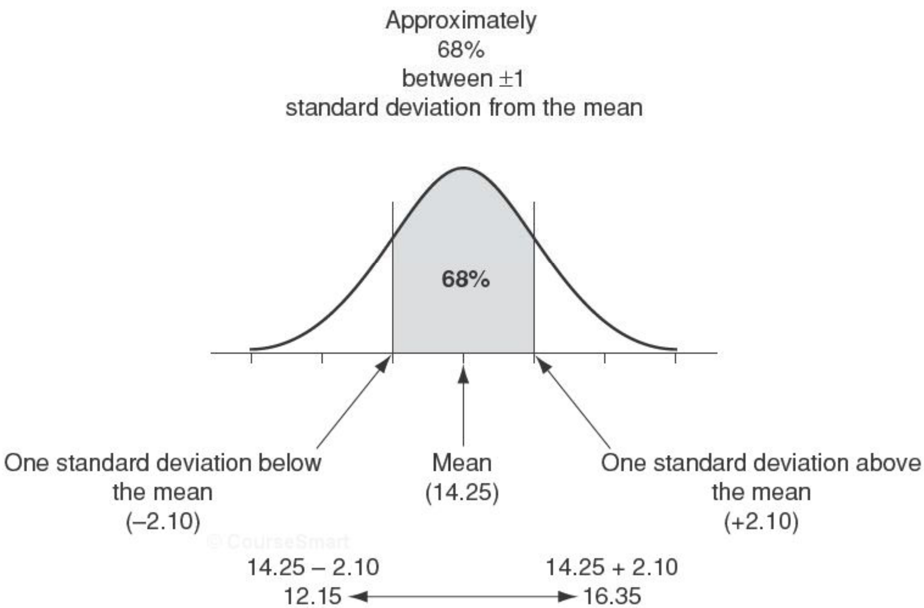


Figure 4-1 Logic Behind the Problem Solution

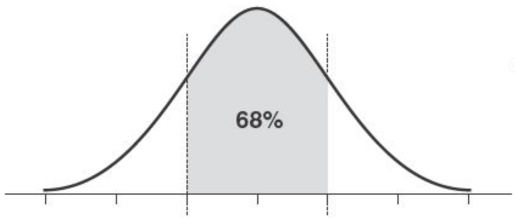
So much for a distribution of data concerning voluntary group counseling. You might study voluntary counseling participation, but another researcher might study the birth weights of a certain type of dog. He/she is apt to discover that the variable of birth weight (like voluntary counseling participation) is normally distributed. Of course, the values of the mean and standard deviation would be different—maybe a mean birth weight of 10.3 ounces with a standard deviation of 1.4 ounces—but the underlying logic would be the same. If you're willing to make a little leap here, you'll no doubt quickly see where we're going with all of this.

One researcher might have normally distributed data measured in hours and minutes, but the next researcher might have normally distributed data measured in pounds and ounces. Someone else might be looking at a variable that is normally distributed and expressed in dollars and cents, while another looks at normally distributed data expressed in years or portions of a year. Different researchers study different variables. It's as simple as that.

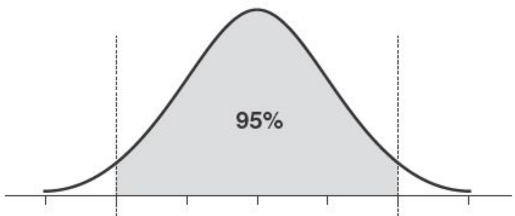
The list could go on and on—an endless array of normal distributions. The different distributions would have different means, different standard deviations, and different underlying scales of measurement (pounds, dollars, years, and so forth), but each normal distribution would conform to the same underlying relationship between the mean and standard deviation of the distribution and the shape of the curve.

The 1-2-3 Rule would always apply: Approximately 68% of the cases would be found ±1 standard deviation from the mean; approximately 95% of the cases would be found ±2 standard deviations from the mean; and more than 99% of the cases would be found ±3 standard deviations from the mean. To review the 1-2-3 Rule, see Figure 4-2.

By the same token, approximately 32% of the cases (or values) under a normal curve would be found *beyond* a value of ±1 standard deviation from the mean. (If approximately 68% of the total area falls within ±1 standard deviation, then the remaining amount—32%—must fall beyond those points.) Similarly, only about 5% of the cases (or values) under a normal curve would be



Approximately 68% of the area under a normal curve is between one standard deviation above and below the mean.



Approximately 95% of the area under a normal curve is between two standard deviations above and below the mean.

More than 99% of the area under a normal curve is between three standard deviations above and below the mean.

Figure 4-2 The 1-2-3 Rule in Review

found *beyond* a point ±2 standard deviations from the mean (100% – 95% = 5%). As for the real extremes of the curve, only about 1% of the area under the curve would be found beyond the points ±3 standard deviations from the mean (100% – 99% = 1%).

Part of what makes the 1-2-3 Rule so useful is the fact that you can use it regardless of the underlying scale of measurement. You know what percentage of scores or values will fall between or beyond certain portions of the curve, regardless of the unit of measurement in question. It doesn't make any difference whether you're dealing with pounds, ounces, dollars, years, or anything else. You know what percentage of cases will be found where—provided the curve is a normal curve. It also doesn't make any difference whether the mean and standard deviation are large numbers (let's say, thousands of dollars) or small numbers (let's say, values between 4 and 15 ounces). Assuming a normal distribution, the 1-2-3 Rule applies. The 1-2-3 Rule is useful because it is expressed in standard deviation units.

So much for normal curves that you're apt to find in real life. Now we come to the matter of the standardized normal curve—a theoretical curve. Let me urge you in advance to be open-minded as we move forward. Indeed, let me caution you not to expect any direct application right away. The applications will come in good time.

Into the Theoretical World

First and foremost, the standardized normal curve is a theoretical curve. It's a theoretical curve because it's based upon an infinite number of cases. Even if you're inclined to move right ahead with the discussion, let me suggest that you take a moment to reflect on that last point: The standardized normal curve is a theoretical curve; it is based on an infinite number of cases.

> **LEARNING CHECK**
>
> *Question:* Why is the standardized normal curve considered a theoretical curve?
>
> *Answer:* It is based on an infinite number of cases.

Here's a way to understand that point. Imagine a normal curve with a line in the middle that indicates the position of the mean. Now envision each side of the curve moving farther and farther out—the right side moving farther to the right and the left side moving farther to the left. Imagine something like the curve shown in Figure 4-3.

Because the standardized normal curve is based on an infinite number of cases, there's never an end to either side of it. As with other normal distributions, the bulk of the cases are found in the center of the distribution (clustered

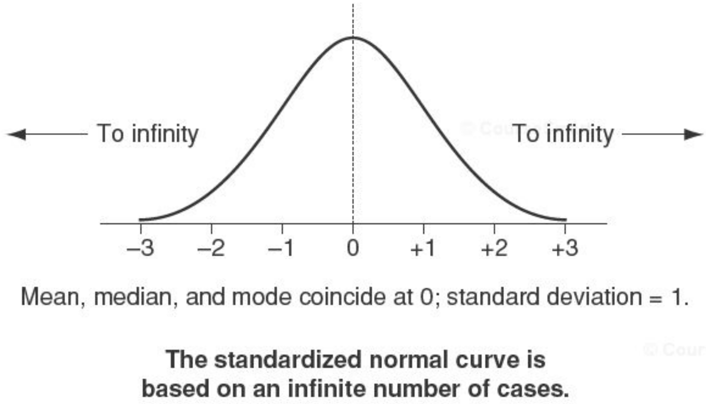


Figure 4-3 Theoretical Nature of the Standardized Normal Curve

around the mean), and the cases trail off from there. As the cases trail off on either side of the distribution, the curve falls ever so gradually toward the baseline. But (and this is an important *but*), the standardized normal curve never touches the baseline. Why? The standardized normal curve never touches the baseline because there are always more cases to consider. (Remember: the curve is based on an infinite number of cases.)

LEARNING CHECK

Question: What is the effect of an infinite number of cases on the curve and the baseline?

Answer: The curve never touches the baseline because there are always more cases to consider.

the standardized normal curve as *the* normal curve. To fully grasp this point, think about the example involving the drug rehabilitation program participants. In that example, the mean was 14.25 hours spent in voluntary group counseling, and the standard deviation was 2.10 hours. You might encounter another normal distribution, though, with a mean of 700 and a standard deviation of 25. At this point, it shouldn't concern you what the 700 and the 25 represent; they could be dollars or pounds or test scores or any number of other variables. The idea is to move your thinking to a more abstract level. Each distribution has a mean and a standard deviation. These values may be expressions of income amounts, test scores, number of tasks completed, growth rates, or any other variable.

In the case of the standardized normal curve, though, the mean is always equal to 0 and the standard deviation is always equal to 1. It is not the case that the mean is, let's say, 16 and the standard deviation is 2. It isn't the case that the mean is 2378 and the standard deviation is 315. You might have means and standard deviations like those in some normal distributions, but what we're considering here is the *standardized* normal curve.

Let me repeat: In the case of the standardized normal curve, the mean is equal to 0 and the standard deviation is 1. These two properties—a mean of 0 and a standard deviation of 1—are the properties that really give rise to the term *standardized*. They're also the properties that make the standardized normal curve so useful in statistical analysis.

We start with the notion that the mean is equal to 0 (see Figure 4-4). Because the mean is equal to 0, any point along the baseline of a normal curve that is above the mean is viewed as a positive value. Likewise, any value below the mean would be a negative value. As you already know, the two sides of any normal curve are equal. Therefore, the area falling between the mean and a certain distance above the mean (on the right side of the curve) is the same as the area between the mean and that same distance on the left side of the curve (below the mean).

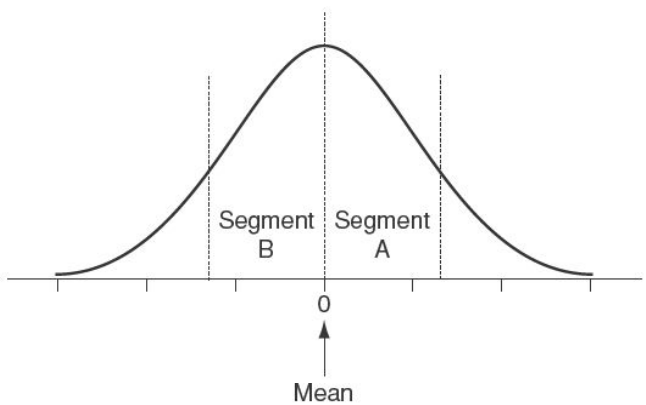


Figure 4-4 Equality of Areas on Both Sides of the Standardized Normal Curve

In a way, the information you just digested cuts your learning in half. The only difference between the two sides of the standardized normal curve is that we refer to points along the baseline as being either positive or negative—positive for points above the mean, and negative for points below the mean. Well and good, but what am I supposed to be learning? you may ask. Patience! We'll get to that. Remember: The idea is to thoroughly digest the information.

The Table of Areas Under the Normal Curve

In a sense, it isn't the standardized normal curve itself that's so useful in statistical analysis. Rather it's the **Table of Areas Under the Normal Curve** that proves to be the really useful tool. You'll find a copy of the Table of Areas Under the Normal Curve in Appendix A, but don't look at it just yet. Instead, follow along with a little more of the discussion first.

To understand just how useful the Table of Areas Under the Normal Curve can be, think back to our previous discussion. Earlier you learned the 1-2-3 Rule, and that gave you some information about areas under a normal curve. But what about areas under the curve that fall, let's say, between the mean and 1.25 standard deviations above the mean? Or what about the area beneath the curve that is found between the mean and 2.17 standard deviations below the mean? In other words, everything is fine if you're dealing with 1, 2, or 3 standard deviations from the mean of a normal distribution, but what about other situations?

With a little bit of calculus, you could deal with all sorts of situations. You could calculate the area under the curve between two points, or the portion under the curve between the mean and any point above or below the mean. Fortunately, though, you don't have to turn to calculus. Thanks to the Table of Areas Under the Normal Curve, the work has already been done for you.

There's a chance that you're muttering something like, What work—what am I supposed to be doing? Relax; lighten up. Remember what the goal is right now—to learn some fundamental material without worrying about its direct application. Concentrate on the basic material right now; the applications will come in due time.

Before I ask you to turn to the Table of Areas Under the Normal Curve (Appendix A), let me say a word about what you're going to encounter and what you'll have to know to make proper use of the table. First, you should take time for a dark room moment to once again imagine what the standardized normal curve looks like. Imagine that you're facing a standardized normal curve. You notice the value of 0 in the middle of the baseline, along with an infinite number of hatch-marks going out to the right and to the left. Also, imagine that the area under the curve is full of cases (just as you did earlier when you were introduced to the notion that the area under the curve isn't just blank space).

Now, instead of thinking about a bunch of hatch-marks that mark points along the baseline, start thinking about the hatch-marks as something called

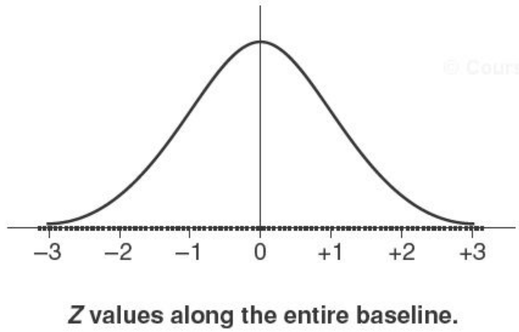


Figure 4-5 Distribution of *Z* Values Along the Baseline of the Standardized Normal Curve

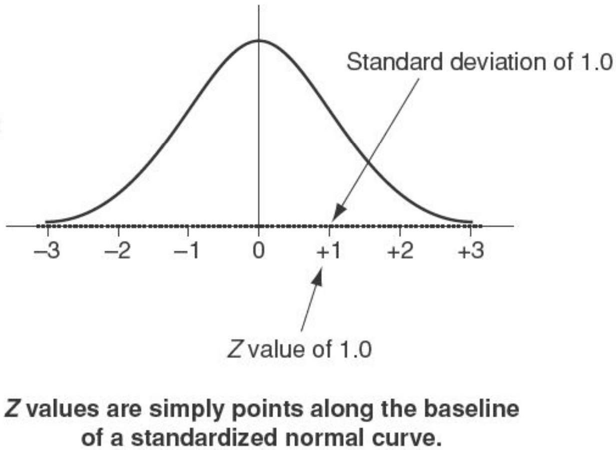


Figure 4-6 *Z* Values as Standardized Deviations Along the Baseline of the Standardized Normal Curve

Z values. The term ***Z***, or ***Z* score**, is the expression statisticians use to refer to points or values along the baseline of the standardized normal curve. The point at the middle of the curve has a *Z* value of 0; other *Z* values are found to the right and to the left of that zero point. The *Z* values on the right are considered positive *Z* values; the *Z* values on the left are considered negative *Z* values (see Figure 4-5).

Since the standard deviation of the standardized normal curve is equal to 1, *Z* values along the baseline are really expressions of standard deviations along the baseline. For example, a *Z* value of +2 really equals 2 standard deviation units above the mean. A *Z* value of –1.3 would equate to 1.3 standard deviation units below the mean. A *Z* value of 0 would be 0 standard deviations away from the mean because it would be equal to the mean. Consider the illustration in Figure 4-6.

Now take a look at Appendix A: Table of Areas Under the Normal Curve. It is also known as *the distribution of Z*. First, focus on the graphs in the illustration on page 308. The illustration lets you know that the table gives you information about the amount of area under the normal curve that's located between the mean and any point along the baseline of the curve. Second, focus on different columns. You'll see the symbol *Z* at the top of several columns. You'll also see columns marked Area Between Mean and *Z*.

The body of the table is filled with proportions (expressed as decimal values). These can easily be translated into percentage values by multiplying by 100. For example, the value of .4922 in the body of the table should be read as 49.22%. The percentage value of 49.22% is associated with a *Z* value of 2.42. How do you know that? Just have a look at the table. The value of .4922 appears next to the *Z* value of 2.42. The best way to understand all of this is to just jump right in and take a look at the table.

Let's say you want to find the proportion or percentage value associated with a *Z* value of 1.86. First you have to locate the *Z* value of 1.86 (see Figure 4-7). Then you look to the right of that *Z* value for the associated proportion. The corresponding proportion value is .4686, which translates into 46.86%. Now you ask, 46.86% of what? Here's the answer: 46.86% of the area under the normal curve is located between the mean and a *Z* value of 1.86. It doesn't make any difference whether it is a *Z* value of +1.86 or a *Z* value of –1.86; the associated proportion (or percentage) value is the same.

Z	Area Between Mean and *Z*	*Z*	Area Between Mean and *Z*	*Z*	Area Between Mean and *Z*	*Z*	Area Between Mean and *Z*
0.00	0.0000	0.50	0.1915	1.00	0.3413	1.50	0.4332
0.01	0.0040	0.51	0.1950	1.01	0.3438	1.51	0.4345
0.29	0.1141	0.79	0.2852	1.29	0.4015	1.79	0.4633
0.30	0.1179	0.80	0.2881	1.30	0.4032	1.80	0.4641
0.31	0.1217	0.81	0.2910	1.31	0.4049	1.81	0.4649
0.32	0.1255	0.82	0.2939	1.32	0.4066	1.82	0.4656
0.33	0.1293	0.83	0.2967	1.33	0.4082	1.83	0.4664
0.34	0.1331	0.84	0.2995	1.34	0.4099	1.84	0.4671
0.35	0.1368	0.85	0.3023	1.35	0.4115	1.85	0.4678
0.36	0.1406	0.86	0.3051	1.36	0.4131	**1.86**	**0.4686**
0.37	0.1443	0.87	0.3078	1.37	0.4147	1.87	0.4693
0.38	0.1480	0.88	0.3106	1.38	0.4162	1.88	0.4699
0.39	0.1517	0.89	0.3133	1.39	0.4177	1.89	0.4706

Z = 1.86

.4686 or 46.86%

Locate the *Z* value of 1.86. The corresponding value (expressed as a proportion) can be converted to a percentage by multiplying by 100. Thus, 46.86% of the area under the normal curve is located between the mean and a *Z* value of 1.86 (either +1.86 or –1.86).

Figure 4-7 A Segment of the Table of Areas Under the Normal Curve

While we're at it, let me point out a couple of things about the table.

1. What you're looking at is simply one format for presenting areas under the normal curve. Different statistics books use different formats to present the same material.
2. Pay attention to the note under the title of the table: Area Between the Mean (0) and *Z*. Think about what that tells you—namely, that the table gives you the amount of area under the curve that will be found between the mean and different *Z* values.
3. Get comfortable with how the values are expressed—as proportions in decimal format. These proportions can easily be converted to percentages. For example, the value of .4686 is the same as 46.86%.
4. You're probably better off if you immediately begin to think of the values in terms of the percentage of cases or observations between the mean and *Z*. In other words, each and every *Z* value has some percentage of cases or observations associated with it.
5. Take note of the end of the table—how it never really gets to a value of .5000 (or 50%). It goes out to a *Z* value of 3.9 (with an associated percentage of 49.99%), but then it ends. That's because the table is based on an infinite number of cases. Note that each time there's a unit change in the *Z* value (as you move further along in the table), the corresponding unit change in the associated area becomes smaller and smaller. That's because the tail of the curve is dropping closer and closer to the baseline as you move further out on the curve.

Now let's start making use of the table—first doing some things to get you familiar with the table, and then making some applications. We'll start with some problems that involve looking up a *Z* value and associated percentage. Always remember that the table only deals with one-half of the area under the curve. Whatever is true on one side of the curve is true on the other—right? Now consider the following questions.

Question: What is the percentage value associated with a *Z* value of +1.12, and how do you interpret that?

Answer: The proportion value is .3686, or 36.86%. This means that 36.86% of the area under the normal curve is found between the mean and a *Z* value of +1.12.

Question: What is the percentage value associated with a *Z* value of –1.50, and how do you interpret that?

Answer: The proportion value is .4332, or 43.32%. This means that 43.32% of the area under the normal curve is found between the mean and a *Z* value of –1.50.

Question: What is the percentage value associated with a *Z* value of +.75, and how do you interpret that?

Answer: The proportion value is .2734, or 27.34%. This means that 27.34% of the area under the normal curve is found between the mean and a *Z* value of +.75.

Question: What is the percentage value associated with a *Z* value of –2.00, and how do you interpret that?

Answer: The percentage value is .4772, or 47.72%. This means that 47.72% of the area under the normal curve is found between the mean and a *Z* value of –2.0.

Question: What is the percentage value associated with a *Z* value of +2.58, and how do you interpret that?

Answer: The proportion value is .4951, or 49.51%. This means that 49.51% of the area under the normal curve is found between the mean and a *Z* value of +2.58.

If you were able to deal with these questions successfully, we can move on to the next few questions. At this point, let me remind you again of what you already know from previous discussions: The table you're working with reflects only one side of the standardized normal curve. Now let's look at some more questions, this time concentrating on the area between two *Z* values.

Question: How much area under the curve is found between the *Z* values of +1.41 and –1.41?

Answer: 84.14% (Double 42.07% to take into account the fact that you're dealing with both sides of the curve.)

Question: How much area under the curve is found between *Z* values of +.78 and –.78?

Answer: 56.46% (Double 28.23% to take into account the fact that you're dealing with both sides of the curve.)

Question: How much area under the curve is found between *Z* values of +1.96 and –1.96?

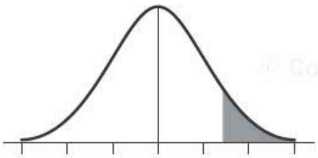
Answer: 95% (Double 47.50% to take into account the fact that you're dealing with both sides of the curve.)

Question: How much area under the curve is found between *Z* values of +2.58 and –2.58?

Answer: 99.02% (Double 49.51% to take into account the fact that you're dealing with both sides of the curve.)

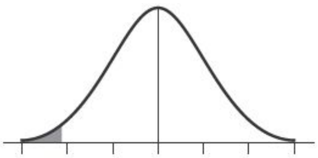
The answers to these questions were fairly straightforward because they simply required that you double a percentage value to get the right answer. You may have already gained enough knowledge about the areas under the normal curve to move forward, but I'd like to make certain that you've developed that second-nature, gut-level understanding that I mentioned earlier. To do that, I'm asking that you consider yet another round of questions.

With any questions about areas under the normal curve, it's usually a good idea to draw a rough diagram to illustrate the question that's being posed. Your



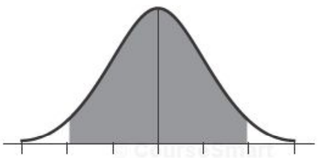
How much area under the curve is above a Z value of +1.44?

Answer: 7.49%.
From mean to Z is 42.51%. Entire half = 50%.
50 – 42.51 = 7.49.



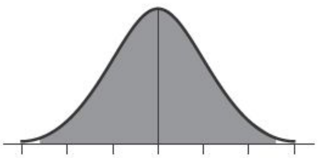
How much area under the curve is below a Z value of –2.13?

Answer: 1.66%.
From mean to Z is 48.34%. Entire half = 50%.
50 – 48.34 = 1.66.



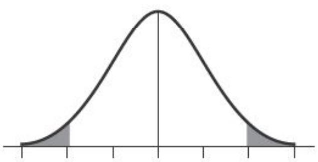
How much area under the curve is between Z values of ±1.96?

Answer: 95%.
From mean to Z is 47.50%. Consider both sides; double the value. 47.50% × 2 = 95%.



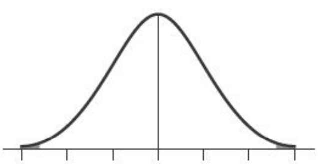
Approximately how much area under the curve is between Z values of ±2.58?

Answer: 99%.
From mean to Z is 49.51%. Consider both sides; double the value. 49.51% × 2 = Approximately 99%.



How much area falls outside of (above and below) the Z values of ±1.96?

Answer: 5%.
The area between Z values of ±1.96 is 95%. The entire area is 100%. 100% – 95% = 5% (evenly split on both sides of the curve).



How much area falls outside of (above and below) the Z values of ±2.58?

Answer: 1%.
The area between Z values of ±2.58 is 99%. The entire area is 100%. 100% – 99% = 1% (evenly split on both sides of the curve).

Figure 4-8 Problems Based on Areas Under the Standardized Normal Curve diagram can be very unsophisticated, just as long as it allows you to put something on paper that expresses the question that's posed and what's going on in your mind when you approach the question. I should warn you to resist any urge to develop your own shortcuts based on the way a question is asked. Always take the time to think through the question. Use a diagram to convince yourself that you're approaching the question the right way.

Now take a look at the questions presented in Figure 4-8, along with the diagrams and commentary. These questions are very similar to some of those you encountered earlier. For these questions, though, focus on how helpful the diagrams are in explaining the underlying logic of the process.

By now you should have noticed that a couple of values have come up time and time again—namely, the values of 95% and approximately 99%. That wasn't by accident. As it turns out, statisticians very often speak in ways that directly or indirectly make reference to 95% or 99%. They're particularly interested in extreme values, cases, or events—the ones that lie beyond the 95% or 99% range. Another way to think of those values is to think of them as being so extreme that they're only apt to occur less than 5 times out of 100 or less than 1 time out of 100. That's why the Z values of ±1.96 and ±2.58 take on a special meaning to statisticians.

As you learned earlier, the area between Z values of ±1.96 on a normal curve or distribution will encompass 95% of the cases or values. Therefore, only 5% of the cases or values fall beyond the Z values of ±1.96 on a standardized normal curve. Similarly, the area between Z values of ±2.58 will take in slightly more than 99% of the cases or values. Therefore, less than 1% of the cases or values are beyond the Z values of ±2.58.

Those areas, the 5% and the 1%, are the areas of the extreme (unlikely) values—and those are the areas that ultimately grab the attention of statisticians. I'll have a lot more to say about that later. Right now, though, my guess is that your patience is running out and you're anxious to get to an application. Wait no longer. We'll move ahead with an example—one that may strike you as strangely familiar.

Finally, an Application

The truth of the matter is that you've already dealt with a partial application of the material you just covered. You did that earlier when you worked through the example in the last chapter involving your test scores. Think back for a moment to what that example involved. Here it is again, repeated just as it was presented earlier:

Test	Mean	Standard Deviation	Your Score
Math	82	6	80
Verbal	75	3	75
Science	60	5	70
Logic	70	7	77

By way of review, here's the situation you encountered earlier: You were part of a fairly large class (200 students); you took four tests; then I asked you some questions about your relative performance on the different tests.

An assumption was made that the distribution of scores on each test was normal. Additionally, the number of cases involved in each test was fairly large (200 cases). In situations like that—situations involving a large number of cases and distributions that are assumed to be normally distributed—you can convert raw scores to Z scores and make use of the Table of Areas Under the Normal Curve.

To understand all of this, think back to how you eventually came to view your performance on the Science Test. The mean on the Science Test was 60, with a standard deviation of 5. Your score was 70. You eventually thought it through and determined that your score was equal to two standard deviation units above the mean. Your score was 10 points above the mean; the standard deviation equaled 5 points; you divided the 10 points by 5 points (the standard deviation). As a result, you determined that your score was equal to two standard deviation units above the mean.

In essence, what you did was convert your raw score to a Z score. Had I introduced the formula for a Z score earlier, it might have caused some confusion or panic. Now the formula should make more sense. Take a look at the formula for a Z score, and think about it in terms of what was going on in your mind as you evaluated your performance on the Science Test.

$$Z = \frac{X - \mu}{\sigma}$$

Don't panic. Just think about what the symbols represent. First, you're dealing with the results of a class of students, and you're making the assumption that the class is a population. In other words, you're dealing with a population, so the mean on the Science Test is labeled μ, or mu. By the same token, the standard deviation is the standard deviation for the population (remember, we're treating the class as a population), so the standard deviation is symbolized by σ. The symbol X represents a raw score—in this case, your test score of 70.

The formula simply directs you to find the difference between a raw score and the mean, and then divide that difference by the standard deviation. For example, here's what was involved in converting your science score (raw score) to a standardized score:

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \\ &= \frac{70 - 60}{5} \\ &= \frac{10}{5} \\ &= 2 \end{aligned}$$

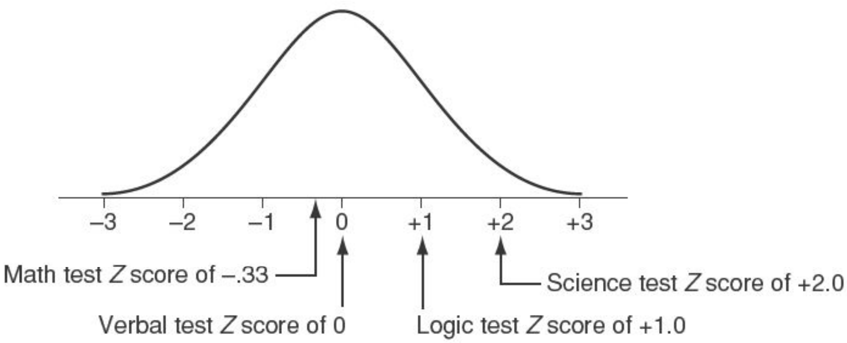


Figure 4-9 Conversion of Test Scores (Raw Scores) to *Z* Scores

When you determined that you scored two standard deviations above the mean, you were simply doing exactly what the formula directs you to do. You found the difference between your score (70) and the mean (60), and you divided that difference by the standard deviation (5). The result is a ***Z* ratio**. It's a ratio of the difference between a raw score and the mean, expressed in standard deviation units.

As shown in Figure 4-9, your score of 70 on the Science Test equated to a *Z* score, or *Z* ratio, of +2. By the same token, your other scores also represented *Z* scores or *Z* ratios.

In each case, you converted your test score to a *Z* score or *Z* ratio by superimposing the distribution of scores for each test onto a single standard—the standardized normal curve. The result was that you could eventually stand back and review all of your test performances in terms of what they were as *Z* scores or *Z* ratios. The results are just the same as they were when the scenario was originally presented in Chapter 2. Your best performance was on the Science Test; your worst performance was on the Math Test.

And just in case you're interested—just in case we had thrown in the Foreign Language Ability Test (like we did in Chapter 2)—it would also have its place on the illustration shown in Figure 4-9. Just to refresh your memory, think back to how the original problem was presented in Chapter 2. There were four 100 point tests to consider. After you had dealt with each of them, you were in a position to determine which was your best and worst performance. But then—at the end and after you thought the matter was settled—I asked you to consider one last scenario. I asked you to consider a situation in which you had also taken a 250 point Foreign Language Ability Test. Additionally, I told you that the mean for the test was 120 with a standard deviation of 15, and I told you that you had scored 90 on the test. If you recall what happened when we did that earlier (i.e., when we added a fifth test to the mix, but it was a 250 point test), then you recall that you had a new *worst performance*. It was the Foreign Language Ability Test score—a score that was two standard deviations below the mean. In short, if we had thrown the Foreign

Language Ability Test into the present scenario, your score on that test would find its rightful place along the baseline shown in Figure 4-9. More specifically, it would be at the point corresponding to a negative Z value—a Z of –2. The Foreign Language Ability Test score would be positioned right where it should be—right there along the baseline, just like the Z values of the other four test scores. Each test would have its own spot along the same baseline—even though four of the tests were 100 point tests and one of the tests (the Foreign Language Ability Test) was a 250 point test.

But there's got to be more to it than just that, you're likely to be saying right now. Truth be known, there is. But patience is called for right now. Remember what the goal is—namely, to develop a solid understanding of the fundamental concepts. For what it's worth, just think about all you've learned so far.

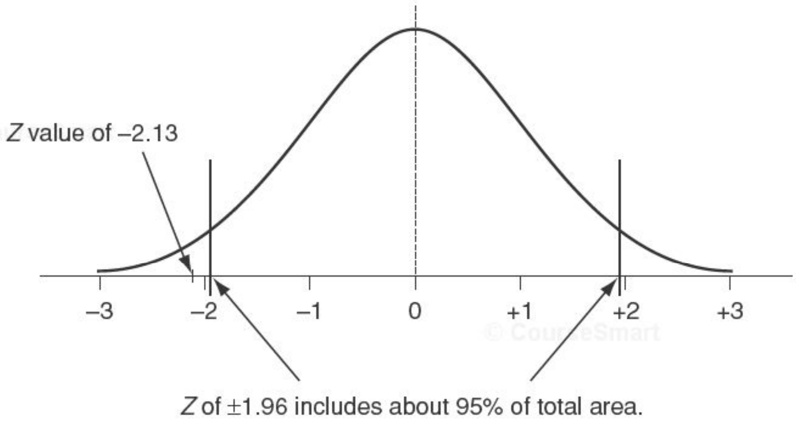
You've made your way through the fundamentals of descriptive statistics and the shapes of distributions in general. What's more, you've just had a solid introduction to the standardized normal curve, Z scores, and the Table of Areas Under the Normal Curve. In the process, you've covered quite a bit.

In learning about the standardized normal curve and The Table of Areas Under the Normal Curve, you've solidified your thinking about curves, distributions, and associated percentages of cases or probabilities of occurrence. More important, you've learned to think in the abstract—assuming you've taken the time to mentally visualize the standardized normal curve and Z scores or points along the baseline. In other words, you've learned to interpret Z scores in a fundamental way.

Let's say, for example, that someone is looking at a raw score value of 62. Then let's say that the 62 equates to a Z value of –2.13. By now, you should automatically know that a Z value of –2.13 is extreme, at least in the sense that it would be located toward the left end of a normal curve. You could look it up on the Table of Areas Under the Normal Curve and find out just how extreme it is, but you should know intuitively that it is extreme. After all, you know that a Z value of –1.96 is extreme, and a Z value of –2.13 would be even more extreme.

If you've committed just a minor amount of information to memory (in this case, the percentage associated with a Z value of ± 1.96), you could say something rather important about that value of –2.13. Without even looking at the Table of Areas Under the Normal Curve, you could make the statement that a value of –2.13 is so extreme that it is likely to occur less than 5 times out of 100 (see Figure 4-10).

Besides automatically knowing the relative position of that Z value, by now you probably have a solid understanding of how the Z value of –2.13 was calculated in the first place. In other words, you understand that the process began by finding the difference between a raw score and the mean of a distribution (in this case, the difference between the mean and 62). That difference was then divided by the standard deviation of the distribution. The result was a Z ratio—a ratio of the difference between a raw score and the mean, expressed in standard deviation units.



Only about 5% of the area would be beyond Z values ±1.96.

In other words, only about 5% of the time would you expect to encounter a Z value that was more extreme than ±1.96. A Z value of –2.13 would be more extreme; therefore, you would expect to encounter it less than 5 times out of 100. In fact, the 5% of the area beyond a Z of ±1.96 would be evenly split, with 2.5% on each side of the curve. Therefore, a value of –2.13 is a value you would expect to occur less than 2.5 times out of 100.

Figure 4-10 Locating a Z Value of –2.13

Now here's the beauty in all of this: It doesn't make any difference whether you're studying weights, heights, incomes, levels of education, levels of aggression in prison inmates, test scores, or anything else. It doesn't make any difference whether you're dealing with values that represent dollars or years or pounds or points or anything else. A Z value (Z ratio) can serve as your standard, just so long as you're dealing with a distribution that has a fairly large number of cases and you can legitimately make the assumption that it is normally distributed.

If your distribution meets those assumptions, you're in a position to know a great deal about your distribution. Most important, you're in a position to identify the extreme values in the distribution. As I mentioned before, it's the extreme values that usually get the attention of statisticians. Indeed, it's usually an extreme result that a statistician is looking at when he/she announces that the results are *significant*.

We'll eventually get into all of that—how to determine whether or not results are statistically significant—but we've still got to cover a few remaining concepts. For that, we go to the next chapter.

Chapter Summary

This chapter was a milestone in the sense that you were introduced to one of the more theoretical but essential concepts in statistical inference—the standardized normal curve. Presumably you learned about the fundamentally theoretical nature of the standardized normal curve, and you learned how to navigate your way around it (with the use of the Table of Areas Under the Normal Curve). What's more, you moved forward on a leap of faith, learning many things about the standardized normal curve with little notion as to where the knowledge would lead.

If the approach worked, though, you eventually found out enough to make your way through some basic applications. Ideally, you moved through those applications with a certain level of intuitive understanding. If that's the way it unfolded for you, welcome to the world of statistical reasoning—you're on the right road. Yes, there are still many more applications to come. But at least you're on the right track.

Beyond that, you were introduced to the fundamental utility of the standardized normal curve—how it allows us to work with a common statistical language, so to speak. You learned that statisticians are typically interested in extreme occurrences. More important, you learned what an extreme occurrence is to a statistician.

I suspect that all of that made for a fairly full plate and a lot to digest at one sitting. Because all that follows is so dependent on what you've just covered, let me urge you to make an honest assessment of your understanding up to this point. If you think you need to reread the material a time or two, make the effort. In many respects, it's one of the keys that unlocks the door.

Some Other Things You Should Know

You deserve to know that the assumption of a normal distribution of a population (or populations, for that matter) is central to many statistical applications. You should also know that it is an assumption that isn't always met. As you might have suspected, statisticians have methods for dealing with situations in which this central assumption cannot be met, but those approaches are beyond the scope of this text. Even if you're eager to learn more about such matters, it pays to remember the old adage, first things first. Since a substantial part of inferential statistics rests on the assumption that you are working with data from a population that is normally distributed, it's essential that you thoroughly cement your understanding of the standardized normal curve.

Beyond that, you should know that there are some relatively easy ways to determine if a distribution is normally distributed—rules of thumb, so to speak, that you can rely upon as a quick alternative to more sophisticated analyses. For example, with a normal distribution, you already know that the mean, median, and mode will coincide. Were you to make a quick check of the values of the mean, median, and mode in a distribution, a substantial difference between or among the values would be an immediate signal that the distribution isn't normal. Similarly, in a normal distribution, you would expect the range divided by 6 to be very close to the value of the standard deviation. Why? You'd expect that because three standard deviations on either side of the mean should take in more than 99% of the area (or cases). Since the mean of a normal distribution would be in the middle of the distribution, you would expect three standard deviations above and below the mean to encompass something close to the total area.

So much for Some Other Things You Should Know at this point. We still have one last bit of information to cover before we really get about the business of inferential statistics, so that's where we'll turn next.

Key Terms

standardized normal curve	*Z* (*Z* score)
Table of Areas Under the Normal Curve	*Z* ratio

Chapter Problems

Fill in the blanks, calculate the requested values, or otherwise supply the correct answer.

General Thought Questions

1. The standardized normal curve is based upon a(n) __________ number of cases.
2. The mean of the normal curve is equal to __________, and the standard deviation is equal to __________.

Application Questions/Problems

1. How much area under the normal curve is between the mean and a *Z* value of 1.63?
2. How much area under the normal curve is between the mean and a *Z* value of 2.35?
3. How much area under the normal curve is between the mean and a *Z* value of –1.22?
4. What percentage of area (cases or observations) is above a *Z* value of +1.96?
5. What percentage of area (cases or observations) is below a *Z* value of –1.96?
6. What percentage of area (cases or observations) is above a *Z* value of +2.58?

7. What percentage of area (cases or observations) is below a Z value of -2.58?
8. What percentage of area under the normal curve is above a Z value of $+1.53$?
9. What percentage of area under the normal curve is below a Z value of -1.12?
10. What Z value corresponds to the lowest 20% of area under the normal curve?
11. What Z value corresponds to the upper 35% of area under the normal curve?
12. What Z values correspond to the middle 60% of area under the normal curve?